Summary of Ridge Regression

Eric Doviak

July 30, 2009

Abstract

This paper – which draws heavily on the work of Robert L. Obenchain – provides an overview of ridge regression. When a researcher finds that the explanatory variables that he or she is working with are highly correlated with each other, the researcher should use ridge regression as a diagnostic for sign stability.

Researchers should consider using ridge regression when working with tax and regulatory variables because public policy variables tend to be highly correlated with each other. The high degree of multicollinearity causes the OLS coefficients to have large standard errors and occasionally causes the estimated coefficients to have the "wrong" sign.

The large standard errors and possible "wrong" signs hamper a researcher's efforts to obtain an estimate of the effect of one variable on a dependent variable of interest, with other variables held constant.

In theory, ridge regression enables the researcher to obtain such an estimate by minimizing MSE risk. In practice, the large number of criteria that the researcher must consider make it difficult to select a particular set of coefficients. Nonetheless, ridge regression provides a useful diagnostic to check for sign stability.

In an extreme case of multicollinearity, the moment matrix (i.e. $X^T X$) of the regression equation (i.e. $\widehat{\beta} = \left(X^T X\right)^{-1} X^T Y$) is exactly singular, so the rank of the matrix is less than the number of regressors. More commonly however, the moment matrix is ill-conditioned, but not of reduced rank.

When working with tax and regulatory variables, it is important to carefully examine the effect of multicollinearity because when two explanatory variables in a regression model are positively correlated, their regression coefficients will be negatively correlated and one of the OLS coefficients might have the "wrong" sign.

A strict interpretation of the OLS coefficients would therefore lead the researcher to conclude that a particular tax or regulatory variable has no effect or the effect opposite of the true effect on the outcome

Table 1: Correlation Matrix

	growth rate	alcohol tax rate	cigarette tax rate
growth rate	1.00		
alcohol tax rate	0.94	1.00	
cigarette tax rate	0.79	0.87	1.00

Table 2: OLS Regression Results

Dependent Variable: Growth Rate		
alcohol tax rate	1.03	***
standard error	0.24	
cigarette tax rate	−0.11	
standard error	0.24	
most likely Q-shape	0.0	
***p-value < 0.01		

of interest. Ridge regression helps diagnose such problems by tracing the paths of the coefficients as they are shrunk towards zero.

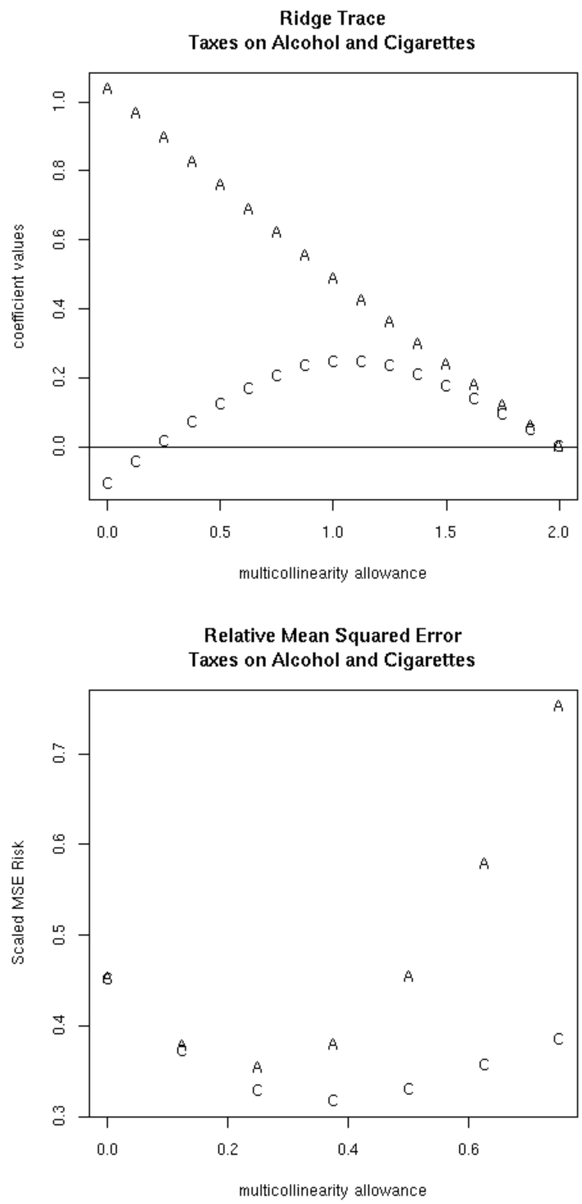
To better understand the effect of multicollinearity, we'll use a fictional dataset that Obenchain provides (2004, chap. 2) and place it in a public policy context. Suppose that the variables in his dataset represent growth rates and cigarette and alcohol tax rates by state. Suppose also that there are strong theoretical reasons to believe that raising tax rates on cigarettes and alcohol increases a state's rate of economic growth.

Given the correlations in Table 1 and our theoretical reasons for believing that cigarette and alcohol taxation accelerate economic growth, one would expect that the coefficients on the tax rates would be positive in a regression on the growth rate. This is not the case however. Table 2 shows that the coefficient on the alcohol tax rate is positive and statistically significant, while the coefficient on the cigarette tax rate is negative and is not statistically significant.

Because the two tax rates are positively correlated with each other, their regression coefficients are negatively correlated with each other. Moreover, the coefficient on the alcohol tax rate should be smaller and the coefficient on the cigarette tax rate should be positive.

Both principal components regression and ridge regression address multicollinearity by effectively reducing the number of dimensions along which the dependent variable is regressed. This produces biased estimates of the full-set of regression coefficients, but can reduce the mean-squared error associated with

Figure 1: Ridge Coefficients and Relative Mean Squared Error



the regression coefficients and correct “wrong” signs.

The difference between ridge regression and principal components arises in the way they reduce the dimensionality. Principal components analysis explicitly reduces the number of dimensions in the regression problem by compressing several variables into one or more composite variables, whereas two-parameter ridge regression effectively reduces the number of dimensions by applying a “shrinkage factor” to each of the regressors. The two-parameter ridge shrinkage factor is:

$$\delta_j = \frac{1}{1 + k\lambda_j^{Q-1}} \tag{1}$$

where λ_j is the eigenvalue associated with parameter j, Q is a "shape parameter" that determines the path that regression coefficients take through likelihood space as they are shrunk toward zero. By definition:

$$0 \leq \delta_j \leq 1. \tag{2}$$

The extent of shrinkage is measured by:

$$M = R - \delta_1 - \delta_2 - \cdots - \delta_R \tag{3}$$

where M is the "multicollinearity allowance" and R is the number of regressors.

Principal components regression is a special case of two-parameter ridge regression because each shrinkage factor is equal to either one or zero when $Q = -\infty$ and $k > 0$.[1] In this special case, integer values are subtracted from the number of regressors and the estimated two-parameter ridge regression coefficients are identical to the principal components regression coefficients (Obenchain, 2004, chap. 3).

Obenchain (1975) shows that the Q-shape most likely to achieve overall minimum MSE risk in estimating regression coefficients is the one which maximizes:

$$COS(Q) = \frac{\sum |r_j{}^o| \lambda_j{}^{(1-Q)/2}}{\sqrt{\sum r_j{}^{o2} \sum \lambda_j{}^{(1-Q)}}} \tag{4}$$

where $r_j{}^o$ is the correlation between the dependent variable and the j-th column of H, the $N \times R$ matrix of standardized principal coordinates of the centered regressor matrix. (H is obtained from the singular value decomposition of the centered regressor matrix). Given the most-likely value of Q, one can – in theory – minimize MSE risk by controlling the extent of shrinkage with the k parameter. In practice however, one must keep other considerations in mind. For example, one would not want to shrink the coefficients so far that they lie outside the 95 percent confidence interval.

In two-parameter ridge regression, the vector of estimated coefficients is given by:

$$\widehat{\beta} = \left(X^T X + k \cdot G \Lambda^Q G^T \right)^{-1} X^T Y \tag{5}$$

where X is the centered regressor matrix, Y is the centered vector of response variables, Λ^Q is an $R \times R$ diagonal matrix of the centered regressor matrix's ordered singular values raised to the Q-th power and G is the $R \times R$ matrix of principal axis direction cosines of the centered regressor matrix. (Like the H matrix, G is also obtained from the singular value decomposition. It's important to note that $GG^T = I$, so that if $Q = 0$ the regression takes the form of Hoerl and Kennard's (1970) classic ridge regression).

Note that when $k = 0$, the estimated coefficients are the ordinary least squares (OLS) estimates.

When working with shrunken estimates from ridge regression, one cannot use conventional t-statistics to test the null hypothesis that the true coefficient value is zero because the expected value of the numerator

[1] There is an exception to this rule. In the rare case that one of the eigenvalues is exactly equal to one, one of the shrinkage factors will be equal to 0.5.

in the t-statistic is not equal to zero (Vinod, 1976). Obenchain (1977) however proves that the OLS t-statistic equals the Ridge t-statistic and argues that practitioners should center their confidence intervals on the OLS estimates.

As an extra diagnostic for our model, it is worthwhile to examine the ridge trace displays to see if the coefficients retain the same sign as they are shrunk towards zero. Some of the coefficients may change sign if the Q-shape parameter is not equal to one. When $Q = 1$, the ridge traces form straight lines because the coefficients have stable relative magnitudes (Obenchain, 2004, chap. 3).

In other cases however, the coefficients do change sign, so before concluding that the coefficient on a tax or regulatory variable is negative and not statistically significant (like the coefficient on cigarette tax rates in the example presented here), we must examine the ridge trace to see if it changes sign as it is shrunk towards zero.

Finally, researchers often want an estimate of the effect of one variable on a dependent variable of interest, with other variables held constant. In principal, the researcher could obtain such an estimate by selecting the set of coefficients that either minimize MSE risk or reduce MSE risk without shrinking the coefficients so far that some of them lie outside of their 95 percent confidence intervals. In practice however, the large number of criteria that the researcher must consider make it difficult to select a particular set of coefficients.

Nonetheless, ridge regression provides a useful diagnostic to check for sign stability. As the tax-growth example shows, the researcher should not conclude that the effect of cigarette taxation on growth is negative after controlling for the effect of alcohol taxation.

References

A. E. Hoerl and R. W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". *Technometrics*, 12(1):55–67, Feb. 1970.

R. L. Obenchain. "Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis". *Technometrics*, 17(4):431–441, Nov. 1975.

R. L. Obenchain. "Classical F-Tests and Confidence Regions for Ridge Regression". *Technometrics*, 19 (4):429–439, Nov. 1977.

R. L. Obenchain. Shrinkage Regression: ridge, BLUP, Bayes, spline & Stein. Mar. 2004. URL `http://members.iquest.net/~softrx/rxadobe.htm`.

H. D. Vinod. "Application of New Ridge Regression Methods to a Study of Bell System Scale Economies". *Journal of the American Statistical Association*, 71(356):835–841, Dec. 1976.